IDIAP RESEARCH REPORT

Recognizing People's Focus of Attention from Head Poses: a Study

Sileye O. Ba [a] Jean-Marc odobez [a]

IDIAP–RR 06-42

July 2006

[a] IDIAP Research Institute

IDIAP Research Institute **www.idiap.ch**
Rue du Simplon 4 P.O. Box 592 1920 Martigny – Switzerland
Tel: +41 27 721 77 11 Fax: +41 27 721 77 12 Email: info@idiap.ch

IDIAP Research Report 06-42

Recognizing People's Focus of Attention from Head Poses: a Study

Sileye O. Ba Jean-Marc odobez

July 2006

Abstract. This paper presents a study on the recognition of the visual focus of attention (VFOA) of meeting participants based on their head pose. Contrary to previous studies on the topic, in our set-up, the potential VFOA of a person is not restricted to other meeting the participants only, but include environmental targets (including a table, a projection screen). This has two consequences. First, it increases the number of possible ambiguities in identifying the VFOA from the head pose. Secondly, in the scenario we present here, full knowledge of the head pointing direction is required to identify the VFOA. An incomplete representation of the head pointing direction (head pan only) will not suffice. In this paper, using a corpus of 8 meetings of 10 minutes average length, featuring 4 persons involved discussing statements projected on a screen, we analyze the above issues by evaluating, through numerical performance measures, the recognition of the VFOA from head pose information obtained either using a magnetic sensor device (the ground truth) or a vision based tracking system (head pose estimates). The results clearly show that in such complex but realistic situations, it is can be optimistic to believe that the recognition of the VFOA can solely be based on the head pose, as some previous studies had suggested.

1 Introduction

The automatic analysis and understanding of human behavior constitutes a rich and interesting research field. It relies on the measurement of the characteristics of one or several people, such as the path they take, their gestures, or their activities (e.g. object handling). One particular characteristic of interest is the *gaze*, which indicates where and what a person is looking at, or, in other words, what is the *visual focus of attention (VFOA)* of the person. In many contexts, identifying the VFOA of a person conveys important information about that person: what is he interested in, what is he doing, how does he explore a new environment or react to different visual stimuli. For instance, tracking the VFOA of people in a public space could be useful to measure the degree of attraction of a given focus target such as advertisements or shop displays. An automatic system based on this principle, such as that presented by [1], would be able to quantify the public exposure of an outdoor advertisement and thus evaluate the adequacy and effectiveness of its content and placement, in a similar manner to claimed recall surveys or empirical traffic studies [2, 3].

Another domain where the gaze plays an important role is human interaction. Indeed, as social being, interacting with other people is an important activity in human daily life, and the way these interactions occur in groups such as families or work teams is the topic of intense study in social psychology [4]. Human interactions happen through speech or non verbal cues. On one hand, the use of verbal cues in groups is rather well defined because it is tightly connected to the taught explicit rules of language (grammar, dialog acts). On the other hand, the usage of the non verbal cues is usually more implicit, which does not prevent it from following rules and exhibiting specific patterns in conversations. A person rising hand often means that he is requesting the floor, and a listener's head nod or shake can be interpreted as agreement or disagreement [5]. Besides hand and head gestures, the VFOA is another important non verbal communication cue with functions such as relationship establishment (through mutual gaze), course of interaction regulation, expressing intimacy, and exercising social control [6]. Speaker's gaze often correlates with the addressees, i.e. the intended recipients of the speech, especially at a sentence end where it can be interpreted as a request of back-channel [7]. Also, on the listener's side, analyzing the speaker's gaze and monitoring his own gaze is a way to find appropriate time windows for speaker turn requests [8, 9]. Furthermore, studies have shown that a person's VFOA was influenced by the VFOA state of other people [6]. Thus, recognizing the visual attention pattern of a group of people can reveal important knowledge about the role of participants, their status such as dominance studied by [10], and the social nature of the occurring interactions. Following psychologists, computer vision researchers are showing more and more interest in the study of automatic gaze and VFOA recognition systems [11, 12, 1], as demonstrated by the tasks defined in several recent evaluation workshops [13, 14]. As an important case, meetings in smart spaces [15], which exemplify the multi-modal nature of human communication and the complex patterns that emerge from interaction, are well suited to conduct such research studies.

In this context, the goal of this paper is to analyze the correspondence between the head pose and the eye gaze of people. In smart spaces such as meeting rooms, it is often claimed that head orientation can be reasonably utilized as an approximation of the gaze [11]. Gaze estimation requires high resolution close-up views, which are generally not available in practice. In this paper, we evaluate the validity of this assumption that gaze can be approximated with head pose by generalizing to more complex situations (VFOA targets requiring the full range of head pose) similar works that have already been conducted by [11] and [12]. Contrary to these previous works, the scenario we consider involves people looking at slides or writing on sheet of paper on the table. As a consequence people have more potential VFOA targets in our set-up (6 instead of 3 in the cited works), leading to more ambiguities between VFOA. Also, due to the physical placement of the VFOA targets, the identification of the VFOA can only be done using the complete head pose representation (pan and tilt), instead of just the head pan as done previously. Thus our study reflects more complex, but realistic, meeting room situations in which people do not just focus their attention on other people but also on other room targets.

To recognize the VFOA of people from their head pose, we adopted a statistical approach. In the static case, each individual pose observation was classified using the Maximum A Posteriori (MAP) classification principle, whereas in the dynamic case, pose observation sequences were segmented into VFOA temporal segments using a Hidden Markov Model (HMM) modeling. In both cases, the head pose observations were represented using VFOA dependent Gaussian distributions.
Alternatives approaches were considered to learn the model parameters. In one approach, a machine learning point of view with training data was exploited. However, as collecting training data can become tedious, we exploit the results of studies on saccadic eye motion modeling [16, 17] and propose another more geometric approach, that models the head pose of a person given his upper body pose and his effective gaze target. This way, no training data are required to learn parameters, but knowledge of the 3D room geometry and camera calibration parameters is necessary. Finally, in practice we observed that people have their own head pose preferences for looking at the same given target. To account for this, we adopted an unsupervised MAP scheme to adapt the parameters obtained from either the learning or geometric model to individual people and meetings.
To assess and evaluate the different aspects of the VFOA modeling (model, parameters, adaptation), we have conducted contrastive and thorough experiments on a significant database that we made publicly available. The database is comprised of 8 meetings of 10-minute average length for which both the head pose ground-truth (captured using magnetic sensor) and VFOA label ground truth are known. Because the head poses are either given by a magnetic sensor (the ground truth) or estimated by a computer vision based probabilistic tracker [18], in our experiments we will be able to differentiate between the two main error sources in VFOA recognition: the use of head pose as proxy for gaze, and errors in the estimation of the head pose.

In summary, the contributions of this paper are the following: 1) the development of a database and a framework to evaluate the recognition of the VFOA solely from head pose 2) a model that provides what a person's head pose should be given her effective gaze target, which exploits prior knowledge about the room geometry and people/target locations 3) the used of an unsupervised MAP framework to adapt the VFOA model parameters to individual meetings, taking into account the specificities of the participants' gaze and the responses of the head pose tracker's , and 4) a thorough experimental study and analysis of the influence of several key aspects on the recognition performance (e.g. participant's position in the meeting room, ground truth vs estimated head pose, correlation with tracking errors).

The remainder of this paper is organized as follows. Section 2 discusses works related to ours. Section 3 describes the task and the database that is used to evaluate the models we propose to solve the task. Section 4 describes the way we obtain head pose measures using either a magnetic field location tracker, or using our probabilistic method for joint head tracking and pose estimation, and compares numerically the latter approach (estimation) with the former (ground truth). Section 5 describes the considered models for recognizing the VFOA from head pose. Section 6 gives the unsupervised MAP framework used to adapt our VFOA model to unseen data. Section 7 describes our evaluation setup. We give experimental results in Section 8 and conclusions in Section 9.

2 Related Work

VFOA is defined by eye gaze, which means the direction toward which the eyes are pointing in the space. Estimating the VFOA requires the ability to detect and track people's eye gaze. Eye gaze tracking methods can be grouped into infrared (wearable) and appearance (non wearable) based tracking methods. In wearable based methods, an infrared light is shined on the subject whose gaze is to be tracked, the difference of reflection between the cornea and the pupil is used to determine the gaze direction which can be used to estimate the VFOA. As an example, [19] studied people's attentions and reactions to advertisement exposure using such technology, to study the best location where an advertiser should put important information to capture clients' attention. However, besides concerns over the safety of long exposure to infrared lights, because of their invasiveness, wearable sensors can

be used only in controlled experimental situations.
In non-controlled situations, non-invasive procedures to estimate the eye gaze are required. This is the case for applications which aim at automatically detecting driver attention loss. In such applications, appearance based eye gaze tracking methods can be used. Appearance based methods, in the presence of high resolution eye images, use image appearance to estimate the gaze direction. [20] use motion and skin color distribution to track a set facial features comprising the eye balls. Gaze direction is reconstructed from the eye ball shape and location. A similar approach was introduced earlier by [21] in the human computer interaction domain to estimate the gaze location of a worker in an office environment. Although gaze tracking is less invasive with computer vision techniques than with a wearable sensor, computer vision gaze tracking techniques are still relatively constraining. The subject has to remain close to the camera because tracking eye features require high resolution images. Thus, [11] proposed to estimate the VFOA using the head pose instead of the eye gaze.
Head pose tracking methods can be categorized into two groups: model based approaches and appearance based approaches. In model based approaches, a set of facial features such as the eyes, the nose and the mouth are tracked. Knowing the relative positions of these features, the head pose is estimated with similar methods such as those proposed by [22] or [23] who give methods to estimate eye gaze and head pose using anthropometric information. Among model-based head pose tracking methods for head pose estimation, we can cite as examples without being exhaustive: [24] who proposed head tracking method which tracked six facial features (eyes, nostrils and mouth borders) and [25] who proposed a stereo-vision based approach. The major drawback of the model based-methods is that they rely on facial feature tracking, which requires high resolution head images. Also, detecting and tracking a small set of feature points is a very difficult task due to occlusions and ambiguities. An alternative to the model-based approach is the appearance-based approach, which does not use specific facial feature, but models the whole head appearance using training data. Because of their robustness in the presence of low resolution head images, appearance based approaches are widely investigated. Among the wide literature on this topic, we can cite as examples: [26, 27, 28] who proposed neural network approaches to model head appearances; [29, 30] who used principal component analysis (PCA) to model head appearances; [31] and [32] who used multidimensional Gaussian distributions to represent the head appearances.
Another perspective for categorizing head pose tracking methods could be "head tracking then pose estimation" versus "joint head tracking and pose estimation". Head tracking then pose estimation consists of tracking the head with a generic tracker to obtain the head location, then extracting features from this location to estimate the head pose [26, 31, 27, 28, 30, 32, 33]. The "head tracking then pose estimation" framework, by decoupling the tracking and head pose estimation process, reduces the configuration space. A smaller configuration space results in a reduction of computational cost. But the relationship between the head spatial configuration and pose is neglected. Knowing the head pose improves head localization and vice versa. Thus, to take into account the mutual relationship between the head spatial configuration and pose, head tracking and pose estimation can be performed jointly, as been done by [29, 25, 18].
When the head pose is available, VFOA can be estimated. As a good example, [11] showed that, in a 4-person meeting configuration, the hypothesis that the head is approximately oriented in the same direction as the gaze is a reasonable assumption. In this work, however, there was no ambiguity between the head poses which indicated people were looking at the VFOA targets, because of the physical meeting set-up (4 participants evenly spaced around a round table). Also, the head poses were reduced to the head azimuth (head pan) only. Following [11], other researchers used the same assumption regarding head pose and eye gaze to model the VFOA of people. For instance [12] make use of the head pan (obtained from a magnetic field head pose tracker sensor) and utterance to infer conversational models in a 4 persons conversation. [34] exploited head pose to model the VFOA in an office and used the VFOA to define workers social geometry (when people are/are not available for communication).

Figure 1: left: the meeting room. right: a sample image of the dataset

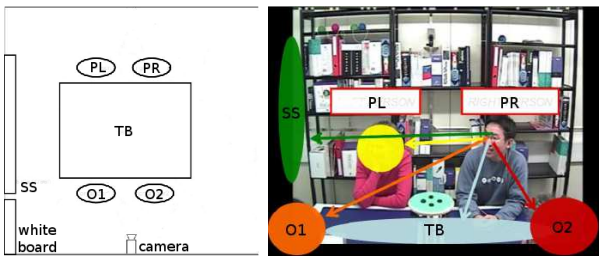


Figure 2: left and right: the VFOA targets.

3 Database and task

In this section, we describe the task and the data that is used to evaluate both our pose estimation algorithm and VFOA recognition algorithm.

3.1 The task and VFOA set

In this work, our goal is to evaluate how well we can infer the VFOA state of a person using head pose in real meeting situations. There are two important issues. The first issue is by definition, the VFOA is given by the eye gaze. However, psycho-visual studies have shown that other cues -e.g. head and body posture, speaking status- play an important role in determining the VFOA state of a person [6]. Thus, the objective is to see, in the absence of gazing information, which may not be available in many applications of interest, how well we can still recognize the VFOA of people. The second issue is the exact definition of a person's VFOA state ? From first thoughts, one can consider that any gaze direction values could correspond to a potential gaze. However, studies about the VFOA in natural conditions [35] have shown that humans tend to look at targets in their environment that are relevant to the task they are solving or of immediate interest to them. Additionally, one interprets another person's gaze not as continuous spatial locations of the 3D space, but as gaze towards objects that has been identified as potential target. This process is often called the shared-attentional mechanism [36, 6]. These studies suggest that VFOA states correspond to a finite set of targets of interests.

Taking into account the above elements, the task is more precisely defined as the following: given the head orientation (the head pose) of a person how to infer his VFOA state. In the context of our meeting set-up and database (see below), the set of potential VFOA targets of interest, denoted $\mathcal{F}$, has been defined as: the other participants to the meeting, the slide-screen, the table, and an additional label (unfocused) when none of the previous could apply. As a person can not focus on himself/herself, the set of focus is thus different from person to person. For instance, for the 'person left' in Figure 2, we have: $\mathcal{F} = \{PR, O2, O1, SS, TB, U\}$ where PR stands for person right, $O1$ and $O2$ for organizer 1 and 2, SS for slide screen, TB for table and U for unfocus. For the person right, we have $\mathcal{F} = \{PL, O2, O1, SS, TB, U\}$, where PL stands for person left.

3.2 The database

Our experiments rely on the IDIAP Head Pose Database[1]. In view of the limitations of visual inspection for evaluation, and the inaccuracy obtained by manually labeling head pose in real videos, we have recorded a video database with head pose ground truth produced by a a magnetic field head orientation tracking sensor. At the same time, in the database, people's discrete VFOA was annotated by hand on the basis of their gaze direction. This allows us to evaluate the impact of using the estimated vs the true head pose as input to the VFOA recognition experiments.

Content description: the database comprises 8 meetings involving 4 people, recorded in a smart meeting room (cf Figure 1, left). The durations of the meetings ranged from 7 to 14 minutes. Our recording were long enough to better represent realistic meeting scenario than short meeting recordings (less than 2 minutes). Because when the meetings are short people use more their head pose to focus to targets. While when the meeting are longer people's attention is sometimes low and they listen without necessarily focusing to speakers.

The scenario for each meeting consisted in writing down one's name on a sheet of, then discussing with the other participants statements displayed on the projection screen. There were restrictions neither on head motions, nor on head poses.

Head pose annotation: in each meeting, the head pose of two persons was continuously annotated (the person left and right in Figure 1 right) using 3D location and orientation magnetic sensors called flock of bird (FOB) rigidly attached to the head, resulting in a video database of 16 different people. The coordinate frame of this sensor was calibrated with respect to the camera frame, and in each recording, the time delay between the FOB and the video was set by detecting the occurrence of the same events (peak oscillations) in both modalities. As a consequence, ground truth of head pose configuration with respect to the camera was generated.
This head pose is defined by three Euler angles (α, β, γ) which parameterize the decomposition of the rotation matrix of the head configuration with respect to the camera frame. Among the possible Euler decompositions, we have selected the one whose rotation axes are rigidly attached to the head to report and comment the results. With this choice we have as can be seen in Figure 3 (Right): the pan angle α representing a head rotation with respect to the y-axis; the tilt angle β representing a head rotation with respect to the x-axis ; and finally the roll angle γ representing a head rotation with respect to the z-axis. Because of the scenario used to record data, people often have negative pan values corresponding to looking at the projection screen. The pan values range from -70 to 60 degree. Tilt values range from -60 (when people are writing) to 15 degrees, and roll value from -30 to 30 degrees.

VFOA annotation: using the predefined VFOA discrete set of targets $\mathcal{F}$, for all the IHPD database, the VFOA of each person (PL and PR) was manually annotated by a single annotator using a multimedia interface. The annotator had access to all data streams, including the central camera view (see Figure 1, left) and the audio. Specific guidance for annotation were defined by [37]. Quality of annotation was evaluated indirectly, on 15 minutes of similar data (same room, same VFOA set). Inter-annotator annotation showed good agreement, with a majority of kappa values higher than 0.8.

4 Head Pose Tracking

Head pose is obtained in two ways: first, from the magnetic sensor readings (see previous Section). This virtually noise-free version is called the ground truth. Secondly, by applying a head pose tracker on the video stream. In this Section, we briefly describe the main components of the computer vision probabilistic tracker that we employed for this purpose. Then, the pose estimation results provided by this tracker are compared with the ground truth and analyzed in detail, allowing ultimately to have a better insight in the VFOA recognition results.

[1] Available at *http://www.idiap.ch/HeadPoseDatabase/* (IHPD)

4.1 Probabilistic Method for Head Pose Tracking

In this subsection, we summarize the Bayesian probabilistic approach described by [18] which was used to track the head pose.
The Bayesian formulation of the tracking problem is well known. Denoting by X_t the hidden state representing the object configuration at time t and by Y_t the observation extracted from the image, the objective is to estimate the filtering distribution $p(X_t|Y_{1:t})$ of the state X_t given the sequence of all the observations $Y_{1:t} = (Y_1, \ldots, Y_t)$ up to the current time. Given standard assumptions (the hidden process is Markovian, and the observations are conditionally independent given the state sequence), Bayesian tracking amounts to solve the following recursive equation:

$$p(X_t|Y_{1:t}) \propto p(Y_t|X_t) \int_{X_{t-1}} p(X_t|X_{t-1}) p(X_{t-1}|Y_{1:t-1}) \mathrm{d}X_{t-1} \tag{1}$$

In non-Gaussian and non linear cases, this can be done recursively using sampling approaches, also known as particle filters. The idea behind particle filter consist in representing the filtering distribution using a set of weighted samples (particles) $\{X_t^n, w_t^n, n = 1, ..., N_s\}$ and updating this representation when new data arrive. Given the particles' set of the previous time step, $\{X_{t-1}^n, w_{t-1}^n, n = 1, ..., N_s\}$, configurations of the current step are drawn from a proposal distribution $X_t \sim \sum_n w_{t-1}^n p(X|X_{t-1})$. The weights are then computed as $w_t \propto p(Y_t|X_t)$. Four elements are important in defining a particle filter:
i) a state model defining the object we are interested in; ii) a dynamical model $p(X_t|X_{t-1})$ governing the temporal evolution of the state; iii) a likelihood model measuring the adequacy of data given the proposed configuration of the tracked object; and iv) a sampling mechanism which have to propose new configurations in high likelihood regions of the state space.
These elements along with our model are described in the next paragraphs.

State Space: The state space contains both continuous and discrete variables. More precisely, the state is defined as $X = (S, \theta, l)$ where S represents the head location and size, θ represents the head in-plane rotation. Both S and θ parameterize a transform $\mathcal{T}_{S,\theta}$ defining the head 2D-spatial configuration. The variable l labels an element of the discretized set of possible out-of-plane head poses[2].

Dynamical Model: The dynamical model governing the temporal evolution of the state is defined as

$$p(X_t|X_{1:t-1}) = p(\theta_t|\theta_{t-1}, l_t) p(l_t|l_{t-1}, S_t) p(S_t|S_{t-1}, S_{t-1}) \tag{2}$$

The dynamics of the head in plane rotation θ_t and discrete head pose l_t variables are learned using head pose GT training data. Head location and size dynamics are modelled as second order auto-regressive processes.

Observation Model: This model $p(Y|X)$ measures the likelihood of the observation for a given state value. The observations $Y = (Y^{text}, Y^{col})$ are composed of texture and color observations (see Fig. 4). Texture features are represented by the output of three filters (a Gaussian and two Gabor filters at different scales) applied at sample locations of the image patch extracted from the image and preprocessed by histogram equalization to reduce light variations effects. Color feature are represented by a binary skin mask extracted using a temporally adapted skin color model. Assuming that, given the state value, texture and color observation are independent, the observation likelihood is modeled as:

$$p(Y|X = (S, \theta, l)) = p_{text}(Y^{text}(S, \theta)|l) p_{col}(Y^{col}(S, \theta)|l) \tag{3}$$

where $p_{col}(\cdot|l)$ and $p_{text}(\cdot|l)$ are pose dependent models. For a given hypothesized configuration X, the parameters (S, θ) allow to extract an image patch, on which the features are computed, while the exemplar index l allows to select the appropriate appearance model.

[2]Note that (θ, l) is another Euler decomposition (using different axis) of the head pose, different than the one described in Subsection 3.2 (cf Figure 3,left). Its main computational advantage is that one of the angle corresponds to the in-plane rotation. It is straightforward to go from one decomposition to the other one.

Figure 3: Left: Training head pose appearance range. Pan angles vary from -90 to 90 degrees with 15 degrees step. Tilt angles vary from -60 to 60 with 15 degrees step. Right: Example of head pose together with it's attached head pose reference; rotation around y-axis represent the head pan, rotation around x-axis represents the head tilt and rotation around z-axis (head pointing direction) represents the head roll

Figure 4: Tracking features. Left: texture features from Gaussian and Gabor filters. Right: skin color binary mask from skin color detection.

Sampling Method: In this work, we use Rao-Blackwellization which consist in applying the standard PF algorithm over the tracking variables S and θ while applying an exact filtering step over the exemplar variable l. The method theoretically results in a reduced estimation variance, as well as a reduction of the number of samples.
For more details about the models and algorithm, the reader is referred to [18].

4.2 Head Pose Tracking Evaluation

For the evaluation of our head pose tracking approach, we followed the protocol described below.

Protocol: We used a two-fold evaluation protocol, where for each fold, we used half (8 people) of our IHPD database (see Sec.3.2) as training set to learn the pose dynamic model and the remaining half as test set.
It is important to note that the pose dependent appearance models were not learned using the same people or head images gathered in the same meeting room environment. We used the Prima-Pointing database [38], which contains 15 individuals recorded over 93 different poses (see Fig. 3). However, when learning appearance models over whole head patches, as done in [18], we experienced tracking failures with 2 out of the 16 people of our evaluation IHPD database (see Section 3) which had hair appearances not represented in the Prima-Pointing dataset (e.g. one of these two persons is bald). As a remedy, we trained the appearance models on patches centered around the visible part *of the face, not the head*. With this modification, no failure was observed, but performance were overall slightly worse than those obtained in [18].

Performance measures: three error measures are used. They are the average errors in pan, tilt and roll angles, i.e. the average over time and meeting of the absolute difference between the pan, tilt and roll of the ground truth (GT) and the tracker estimation. Additional statistics are also given, such as

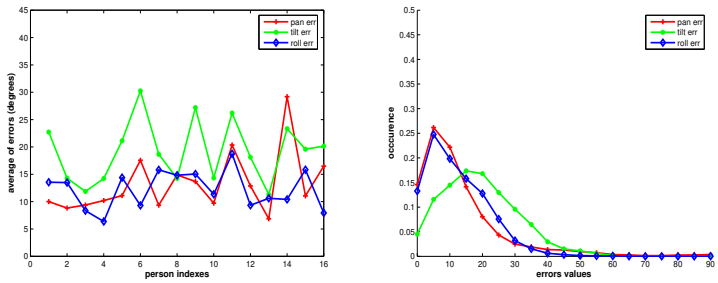


Figure 5: Pan, tilt and roll tracking errors. Left: average errors for each person (R for right and L for left person). Right: distribution of tracking errors over the whole dataset.

side	right persons			left persons		
stat	mean	std	med	mean	std	med
pan	11.4	11	8.9	14.9	12.4	11.3
tilt	19.8	9.1	19.4	18.6	10.3	17.1
roll	14	9.2	13.2	10.3	7.7	8.7

Table 1: global pan, tilt, and roll error statistics for right and left persons.

the median value, which is less affected by large errors which can be due to erroneous tracking.

Results: The statistics of the errors are shown in Table 1. Overall, given the small head size, and the fact that the appearance training set is composed of faces recorded in an external set up (different people, different viewing conditions), the results are quite good, with a majority of head pan errors smaller than 12 degrees (see Figure 5). However these results hide a large discrepancy between individuals. For instance, the average pan error ranges from 7 degrees to 30 degrees, and depends mainly on whether the tracked person's appearance is well represented by the appearances of persons present in the training set used to learn the appearance model. The Table 1 also shows that the pan and roll tracking errors are smaller than the tilt errors. The main reason is that tilt estimation is more dependent on face localization and individual face feature distances than head pan and roll, as pointed out by other researchers [32]. Indeed, even from a perceptive point of view, discriminating between head tilts is more difficult than discriminating between head pan or head roll.

Table 2 details the errors depending on when people true pose are near frontal pose or near profile . We can observe that when the head pose is near the frontal position (pan $|\alpha| \leq 45$ degrees or tilt $|\beta| \leq 30$ degrees), the head pose tracking estimates are more accurate, in particular for the pan and roll value. This can be understood since near profile poses, a pan variation introduces much less appearance changes than the same pan variation near a frontal view. Similarly, for high tilt values, the face-image distortion introduced by the extreme rotation affects the quality of the observations.

Finally, these results are comparable to those obtained by others in similar conditions. For instance, [31] achieved a pan estimation error of 19.2 degrees when true head poses were near pan frontal and 16.9 degrees of pan estimation errors for near pan profile. In another work by [11], a neural net is used to train a head pose classifier from data recorded directly in two meeting rooms. When using 15 people for training and 2 for testing, average errors of 5 degrees in pan and tilt are reported. However, when training the models in one room and testing on data from the other meeting room, the average errors rise to 10 degrees. This suggests an appearance model fitted to the set-up, to the contrary of our experiments, in which appearance models are trained from an external database.

range	pan near frontal			pan near profile			tilt near frontal			tilt far from frontal		
statistic	mean	std	med	mean	std	med	mean	std	med	mean	std	med
pan	11.6	9.8	9.5	16.9	11.1	14.7	12.7	10.4	10	18.6	12.6	15.9
tilt	19.7	9.2	18.9	17.5	8.1	17.5	19	8.8	18.8	22.1	9.2	21.4
roll	10.1	7.6	8.8	18.3	6.2	18.1	11.7	7.7	10.8	18.1	10.6	16.8

Table 2: pan, tilt, and roll error statistics for right person when the true head pan is near frontal (pan $|\alpha| \leq 45$ degrees) or near profile ($|pan| > 45$ degrees) and when the true head tilt is near frontal (tilt $|\beta| \leq 30$ degrees) or not ($|\beta| > 30$ degrees)

5 Visual Focus of Attention Modeling

In this Section, we first describe the models used to recognize the VFOA from the head pose measurements, then the two alternatives we adopted to learn the model parameters.

5.1 Modeling VFOA with a Gaussian Mixture Model (GMM)

Let $s_t \in \mathcal{F}$ and z_t respectively denote the VFOA state and the head pointing direction of a person at a given time instant t. The head pointing direction is defined by the head pan and tilt angles, i.e. $z_t = (\alpha_t, \beta_t)$, since the head roll has no effect on the head direction (see Figure 3). Estimating the VFOA can be posed in a probabilistic framework as finding the VFOA state maximizing the a posteriori probability:

$$\hat{s}_t = \arg\max_{s_t \in \mathcal{F}} p(s_t|z_t) \text{ with } p(s_t|z_t) = \frac{p(z_t|s_t)p(s_t)}{p(z_t)} \propto p(z_t|s_t)p(s_t) \tag{4}$$

For each possible VFOA $f_i \in \mathcal{F}$ which is not *unfocus*, $p(z_t|s_t = f_i)$ is modeled as a Gaussian distribution $\mathcal{N}(z_t; \mu_i, \Sigma_i)$ with mean μ_i and full covariance matrix Σ_i. Besides, $p(z_t|s_t = unfocus) = u$ is modeled as a uniform distribution. We defined the unfocused uniform probability value as $u = \frac{1}{180 \times 180}$ as the head pan and tilt angle can vary from -90 to 90 degrees. In Equation 4, $p(s_t = f_i) = \pi_i$ denotes the prior information we have on the VFOA target f_i. Thus, in this modeling, the pose distribution is represented as a Gaussian Mixture Model (plus one uniform mixture), with the mixture index denoting the focus target:

$$p(z_t|\lambda_G) = \sum_{s_t} p(z_t, s_t|\lambda_G) = \sum_{s_t} p(z_t|s_t, \lambda_G)p(s_t|\lambda_G) = \sum_{i=1}^{K-1} \pi_i \mathcal{N}(z_t; \mu_i, \Sigma_i) + \pi_K u \tag{5}$$

where $\lambda_G = \{\mu = (\mu_i)_{i=1:K-1}, \Sigma = (\Sigma_i)_{i=1:K-1}, \pi = (\pi_i)_{i=1:K}\}$ represents the parameter set of the GMM model. Figure 12 shows is splitted the pan-tilt space according the the VFOA GMM distribution.

5.2 Modeling VFOA with a Hidden Markov Model (HMM)

The GMM approach does not account for the temporal dependencies between the VFOA events. To introduce such dependencies, we consider the Hidden Markov Model. Denoting by $s_{0:T}$ the VFOA sequence, and by $z_{1:T}$ the observation sequence, the joint posterior probability density function of states and observations can be written:

$$p(s_{0:T}, z_{1:T}) = p(s_0) \prod_{t=1}^{T} p(z_t|s_t)p(s_t|s_{t-1}) \tag{6}$$

In this equation, the emission probabilities $p(z_t|s_t = f_i)$ expressing the likelihood of the pose observations for a given VFOA state are modeled as in the previous case (i.e. Gaussian distributions for regular VFOA, and uniform distribution for the *unfocus* VFOA). However, in the HMM modeling, the

static prior distribution on the VFOA targets is replaced by a discrete transition matrix $A = (a_{i,j})$, defined by $a_{i,j} = p(s_t = f_j | s_{t-1} = f_i)$, which models the probability of passing from a focus f_i to a focus f_j. Thus, the set of parameters of the HMM model is $\lambda_H = \{\mu, \Sigma, A = (a_{i,j})_{i,j=1:K}\}$.
With this model, given the observations sequence, the VFOA recognition is done by estimating the optimal sequence of VFOA which maximizes $p(s_{0:T}|z_{1:T})$. This optimization is efficiently conducted using the Viterbi algorithm [39].

5.3 Parameter Learning using Training Data

Since in many meeting settings, people are most of the time static and seated at the same physical positions, setting the model parameters can be done by using a traditional machine learning approach which assumes the availability of training data. Thus, given such data sequences where the VFOA have been annotated, and the head pose measurements have been extracted, we can readily estimate all the parameters of the GMM or HMM models. Parameters learnt with this training approach will be denoted with a l superscript. Note that μ_i^l and Σ_i^l are learnt by first computing the VFOA means and covariances per meeting and then averaging the obtained results on the meetings belonging to the training set.

Prior Distribution and Transition Matrix: While the estimation of the Gaussian parameters using the training data seems appropriate, learning the VFOA prior distribution π or transition matrix A using the annotated data can be problematic. If the training data exhibit specific meeting structure, as it is the case in our database where the main and secondary organizers always occupy the same seats, the learned prior will sometimes have a boosting effect on the recognition performances for similar unseen meetings, boosting effect that we will observe in our experiments. However, at the same time, this learned prior can considerably limit the generalization to other data sets, since by simply exchanging seats of participants having different roles, we can obtain meeting sessions with very different prior distributions. Thus, we investigated alternatives that avoid favoring any meeting structures. In the GMM case, this was done by considering a uniform distribution (denoted π^u) over the prior π. In the HMM case, the transition matrix was designed to exhibit a uniform stationary distribution. Self-transitions defining the probability of keeping the same focus were favored, but transitions to other focus were distributed uniformly according to: $a_{i,i} = \epsilon < 1$, and $a_{i,j} = \frac{1-\epsilon}{K-1}$ for $i \neq j$. Depending on the ϵ value, keeping the same focus is more or less favored. We will denote by A^u the transition matrix built this way.

5.4 Parameter Learning using a Geometric Model

The training approach to parameter learning is straightforward to apply when annotated data is available. However, annotating the VFOA of people in video recording is tedious and time consuming, as training data needs to be gathered and annotated for each (location, VFOA target) couple, the number of which can grow quickly, especially if some moving people are involved. Thus, to avoid the need for annotation, one can seek for an alternative approach that exploits the geometric nature of the problem. The parameters set with the geometric approach described below will be denoted with a superscript g (e.g. μ_i^g).

Assuming a calibrated camera w.r.t. to the room, given a head location and a VFOA target location, it is possible to derive the Euler angles (w.r.t. the camera) for which the head is oriented toward the VFOA target. However, gazing at a target is usually accomplished by moving both the eyes ('eye-in-head' rotation) and the head in the same direction. Researchers working on this topic have found that the relative contribution of the head and eyes towards a given gaze shift follows simple rules [16, 35]. While the experiments conducted in these papers do not completely match the meeting room scenario, we have exploited them to propose a model for predicting a person's head pose given his gaze target.

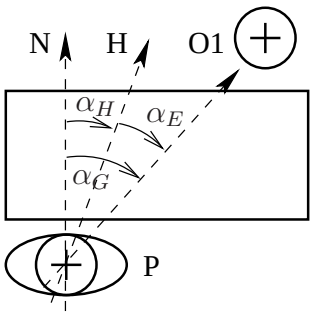


Figure 6: Model of gazing and head orientation.

The proposed geometric model is presented in Figure 6. Given a person P whose rest head pose corresponds to looking straight ahead in the N direction, and given that she is gazing towards O1, the head points in direction H according to:

$$\alpha_H = \kappa_\alpha \, \alpha_G \qquad \text{if } |\alpha_G| < \xi_\alpha, \qquad 0 \text{ otherwise} \tag{7}$$

where α_G and α_H denotes respectively the pan angle to look at the gaze target and the actual pan angle of the head pose. The parameters of this model, κ_α and ξ_α, are constants independent of the VFOA gaze target, but usually depend on individuals [16]. While there is a consensus among researchers about the linearity aspect of the relation between the gaze direction and the head pose direction described by Equation 7, some researchers reported observing head movements for all VFOA gaze shift amplitude (i.e. ξ_α=0), while others do not. In this paper, we will assume $\xi_\alpha = 0$. Besides, Equation 7 is only valid if the contribution of the eyes to the gaze shift (given by $\alpha_E = \alpha_G - \alpha_H$) do not exceed a threshold, usually taken at $\sim 35^o$. Finally, in [16], it is shown that the tilt angle follows a similar linearity rule. However, in this case, the contribution of the head to the gaze shift is usually much lower than for the pan case. Typical values range from 0.2 to 0.3 for κ_β, and 0.5 to 0.8 for κ_α.

In the experiments, we will test the use of this geometric model to predict the mean angles μ in the VFOA modeling. As for the rest reference direction N (Fig 6), we will assume that for the people seated at the two tested positions, it corresponds to looking straight in front of them. Thus, for person left (resp. right), N consists of looking at organizer 1 (resp. 2), as shown in Figure 2. The covariances Σ of the Gaussian distributions will be set according to the size of the targets (i.e. same covariance for each of the 3 meeting participants, and larger for the slide-screen and the table). Finally, the parameter setting of the prior will follow the same considerations than in the previous subsection.

6 VFOA Models Adaptation

In the previous Section, we proposed two models (GMM and HMM) to recognize the VFOA of people from their head pose, along with two approaches to learn their VFOA target dependent parameters: one relying on given training data, and one on information on the room's geometry. Thus, the models we obtained are generic and can be applied indifferently to any new person seated at the location related to a learned model.

However, in practice, we observed that people have personal ways of looking to targets (see Figure 7). For example, some people use less their eye-in-head rotation capabilities and orient more their head towards the focused target than others. In addition, our head pose tracking system is sensitive to the appearance of people, and can introduce a systematic bias in the estimated head pose for a given person, specially in the estimated head tilt.

As a consequence, the parameters of the generic models might not be the best representation for a given person. As a remedy we propose to exploit the Maximum A Posteriori (MAP) estimation

Figure 7: People personal ways of looking: in the two images the two PR are looking to the same target O1 with using different head poses.

principle to adapt, in an unsupervised fashion, the generic VFOA models to the data of each new meeting, and thus produce models adapted to individual person's characteristics.

6.1 VFOA Maximum a Posteriori (MAP) Adaptation

The MAP adaptation principle is the following. Let $z = z_1 ..., z_T$ denotes a set of T samples (i.i.d or drawn from a Markov chain), and $\lambda \in \Lambda$ the parameter vector to be estimated from these sample data. The MAP estimate $\hat{\lambda}$ of the parameters is then defined as:

$$\hat{\lambda} = \arg\max_{\lambda \in \Lambda} p(\lambda|z) = \arg\max_{\lambda \in \Lambda} p(z|\lambda)p(\lambda) \tag{8}$$

where $p(z|\lambda)$ is the data likelihood model which generates the sequence of samples and $p(\lambda)$ is the prior we have on the parameters. If λ is assumed to be fixed but unknown, than this is equivalent to having a non-informative prior $p(\lambda)$, and the MAP estimate reduces to the maximum likelihood (ML) estimate.
The choice of the prior distribution is crucial for the MAP estimation. [40] showed that if the maximum likelihood parameter estimation of the data likelihood model $p(z|\lambda)$ can be conducted using the Expectation-Maximization (EM) algorithm, then, by selecting the prior pdf on λ as the product of appropriate conjugate distributions of the data likelihood[3], then the MAP estimation can also be solved using the EM algorithm. In the next two Subsections, we describe with more details the adaptation equations for our GMM and HMM VFOA models.

6.2 GMM MAP Adaptation

In the GMM VFOA model case, the data likelihood is $p(z|\lambda_G) = \prod_{t=1}^{T} p(z_t|\lambda_G)$, where $p(z_t|\lambda_G)$ is the mixture model given in Equation 5, and λ_G are the parameters to learn which comprise the multinomial prior distribution on the VFOA indices π and the Gaussian parameters of the mixture components μ and Σ. The reader should notice that the presence of a uniform distribution as one mixture does not modify the GMM MAP adaptation framework.

Priors distribution on parameters. For this model, there does not exist a joint conjugate prior density for the parameters λ_G. However, it is possible to express the prior probability as a product of individual conjugate priors [40]. Accordingly, the conjugate prior of the multinomial mixture weights is the Dirichlet distribution $\mathcal{D}(sw_1, \ldots, sw_K)$ whose density function is given by:

$$p^{\mathcal{D}}_{sw_1,\ldots,sw_K}(\pi_1, \ldots, \pi_K) \propto \prod_{k=1}^{K} \pi_i^{sw_i - 1} \tag{9}$$

[3] A prior distribution $g(\lambda)$ is the conjugate distribution of a likelihood function $f(z|\lambda)$ if the posterior $f(z|\lambda)g(\lambda)$ belongs to the same distribution family than g.

- Initialization of $\hat{\lambda_G}$: $\hat{\pi}_i = w_i,\ \hat{\mu}_i = m_i,\ \hat{\Sigma}_i = V_i/(\alpha - p)$
- EM: repeat until convergence:
  1. Expectation: compute c_{it} as well as $\bar{z}_i$ and S_i (Equations 12 and 13) using the current parameter set $\hat{\lambda_G}$
  2. Maximization: update parameter set $\hat{\lambda_G}$ using the re-estimations formulas (Equations 14-16)

Figure 8: GMM adaptation algorithm iterations

Additionally, the conjugate prior for the Gaussian mean and covariance matrix inverse of a given mixture is the Normal-Wishart distribution, $\mathcal{W}(\tau, m_i, \alpha, V_i)$ $(i = 1, ..., K-1)$, with density function

$$p_i^{\mathcal{W}}(\mu_i, \Sigma_i^{-1}) \propto |\Sigma_i^{-1}|^{\frac{\alpha-p}{2}} \exp\left(-\frac{\tau}{2}(\mu_i - m_i)'\Sigma_i^{-1}(\mu_i - m_i)\right) \times \exp(-\frac{1}{2}tr(V_i\Sigma_i^{-1})),\ \alpha > p \tag{10}$$

where $(\mu_i - m_i)'$ denotes the transpose of $(\mu_i - m_i)$, and p denotes the samples' dimension (in our case, $p = 2$). Thus the prior distribution on the set of all the parameters is defined as

$$p(\theta) = p^{\mathcal{D}}_{sw_1,...,sw_K}(\pi_1, \ldots, \pi_K) \prod_{i=1}^{K-1} p_i^{\mathcal{W}}(\mu_i, \Sigma_i^{-1}) \tag{11}$$

EM MAP Estimate. The MAP estimate $\hat{\lambda_G}$ of the distribution $p(z|\lambda_G)p(\lambda_G)$ can be computed using the EM algorithm by recursively applying the following computations (see Figure 8) from [40]:

$$c_{it} = \frac{\hat{\pi}_i p(z_t|\hat{\mu_i}, \hat{\Sigma}_i)}{\sum_{j=1}^{K} \hat{\pi}_j p(z_t|\hat{\mu_j}, \hat{\Sigma}_j)} \text{ and } c_i = \sum_{t=1}^{T} c_{it} \tag{12}$$

$$\bar{z}_i = \frac{1}{c_i}\sum_{t=1}^{T} c_{it} z_t \text{ and } S_i = \frac{1}{c_i}\sum_{t=1}^{T} c_{it}(z_t - \bar{z}_i)(z_t - \bar{z}_i)' \tag{13}$$

where $\hat{\lambda_G} = (\hat{\pi}, (\hat{\mu}, \hat{\Sigma}))$ denotes the current parameter fit. Given these coefficients, the M step re-estimation formulas are given by:

$$\hat{\pi}_i = \frac{sw_i - 1 + c_i}{s - K + T} \tag{14}$$

$$\hat{\mu}_i = \frac{\tau m_i + c_i \bar{z}_i}{\tau + c_i} \tag{15}$$

$$\hat{\Sigma}_i = \frac{V_i + c_i S_i + \frac{c_i \tau}{c_i + \tau}(m_i - \bar{z}_i)(m_i - \bar{z}_i)'}{\alpha - p + c_i} \tag{16}$$

For the uniform component $(i = K)$, the appropriate uniform distribution is used in c_{it} (i.e $p(z_t|\hat{\mu_K}, \hat{\Sigma}_K)$ is indeed a uniform density), and, accordingly, only the prior weight π_K needs to be updated. The choice of the hyper-parameters of the prior distribution $p(\lambda_G)$ in Equation 11 is important as the adaptation is unsupervised. Essentially, only the prior distribution prevents the adaptation process to deviate from meaningful VFOA distributions. The hyper-parameter setting is discussed at the end of this Section.

6.3 VFOA MAP HMM Adaptation

The VFOA HMM can also be adapted in an unsupervised way to new test data using the MAP framework [40]. The parameters to adapt in this case are the transition matrix and the emission probabilities parameters $\lambda_H = \{A, (\mu, \Sigma)\}$[4].
The adaptation of the HMM parameters leads to a procedure similar to the GMM adaptation case. Indeed, the prior on the Gaussian parameters follows the same Normal-Wishart density (Equation 10), and the Dirichlet prior on the static VFOA prior is replaced by a Dirichlet prior on each row $p(.|s = f_i)$ of the transition matrix. Accordingly, the full prior is:

$$p(\lambda_H) \propto \prod_{i=1}^{K} p^{\mathcal{D}}_{sb_{i,1},\ldots,sb_{i,K}}(a_{i,1}, \ldots, a_{i,K}) \prod_{i=1}^{K-1} p_i^{\mathcal{W}}(\mu_i, \Sigma_i^{-1}) \tag{17}$$

Then the EM algorithm to compute the MAP estimate can be conducted as follows. For a sequence of observations, $z = (z_1, ..., z_T)$, the hidden states are now composed of a corresponding sequence of states $s_1, .., s_T$, which allows to compute the joint state-observation density (cf Equation 6). Thus, in the expectation step, one need to compute both $\xi_{i,j,t} = p(s_{t-1} = f_i, s_t = f_j|z, \hat{\lambda_H})$ and $c_{i,t} = p(s_t = f_i|z, \hat{\lambda_H})$, which respectively denote the expected transition probability from state f_i to state f_j at time t and the probability of being in state f_i at time t, given the current model $\hat{\lambda_H}$ and the observed sequence z. These values can be obtained using the Baum-Welch forward-backward algorithm [39]. Given these values, the re-estimation formulas for the mean and covariance matrices are the same as in Equation. 14-16, while the adaptation formula for the transition matrix parameters is given by:

$$\hat{a}_{i,j} = \frac{sb_{i,j} - 1 + \sum_{t=1}^{T-1} \xi_{i,j,t}}{s - K + \sum_{j=1}^{K} \sum_{t=1}^{T-1} \xi_{i,j,t}}. \tag{18}$$

The discussion about how to select the hyper-parameters is conducted in the the following.

6.4 Choice of Prior Distribution Parameters

In this Section we discuss the impact of the hyper-parameter setting on the MAP estimates, through the analysis of the re-estimation formula (Equation 14-16). Before going into details, recall that T denotes the size of the data set available for adaptation, and K is the number of VFOA targets, i.e. the number of GMM states.
Parameter values for the Dirichlet distribution: The Dirichlet distribution modeling the prior on the mixture weights, is defined by two kind of parameters: a scale factor s and the prior values on the mixture weights w_i (with $\sum_i w_i = 1$). The scale factor s controls the balance between the mixture prior distribution w and the data. If s is small (resp. large) with respect to $T - K$, the adaptation is dominated by the data (resp. the prior, i.e. almost no adaptation occurs). When s=$T - K$, data and prior contribute equally to the adaptation process. In the experiments, the hyper-parameter s will be selected through cross-validation among the values in $C^s = \{s_1 = T-K,\ s_2 = 2(T-K)\ , s_3 = 3(T-K)\}$. The priors weights w_i, on the other hand, are defined according to the prior knowledge we have on the VFOA targets distribution. More likely VFOA targets such as the person who speak the most or the slide screen should be given a higher weight. When we want to enforce no knowledge about the VFOA targets distribution, the w_i can be set uniformly equal to $\frac{1}{K}$.
Parameter values for the Normal-Wishart distribution: This distribution defines the prior on the mean μ_i and covariance Σ_i of one Gaussian. The adaptation of the mean is essentially controlled by two parameters (see Equation. 14): the prior value for the mean, m_i, which will be set to the values computed either using a learning (μ_i^l, cf Subsection 5.3) or a geometric approach (μ_i^g cf Subsection 5.4), and a scalar τ, which linearly controls the contribution of the prior value m_i and the data mean $\bar{z}_i$ to the estimated mean. As the average value for c_i is $\frac{T}{K}$, in the experiments, we will select τ though

[4]For convenience, we assumed that the initial state distribution followed a uniform distribution.

cross-validation among the values in $C^\tau = \{\tau_1 = \frac{T}{2K},\ \tau_2 = \frac{T}{K},\ \tau_3 = \frac{2T}{K},\ \tau_4 = \frac{5T}{K}\}$. Thus, with the first value τ_1, the mean adaptation is on average dominated by data. With τ_2, the adaptation is balanced between data and prior, and with the two last values, adaptation is dominated by the priors on the means.
The prior on the covariance is more difficult to set. It is defined by the Wishart distribution parameters, namely the prior covariance matrix V_i and the number of degree of freedom α. However, from Equation 16, we see that the data covariance and the deviation of the data mean from the mean prior also influence the MAP covariance estimate. As prior Wishart covariance, we will take $V_i = (\alpha - p)\tilde{V}_i$, where $\tilde{V}_i$ is respectively either Σ_i^l or Σ_i^g, the covariance of target f_i estimated using either labelled training data (Subsection 5.3) or the geometrical VFOA target size (Subsection 5.4). The weighting $(\alpha - p)$ is important, as it allows V_i to be of the same order of magnitude than the data variance $c_i S_i$, as far as c_i and $(\alpha - p)$ are of similar order of magnitude as well. In the experiments, we will use $\alpha = \frac{5T}{K}$, which put emphasis on the prior, and allow adaptation that do not deviate too much from the covariance priors.

7 Evaluation Set Up

The evaluation of the VFOA models presented previously was conducted using the IHPD database presented in Section 3. Below, we first describe the performance measures we propose to evaluate the VFOA recognition, then give details about the protocols we followed in the experiments.

7.1 Performance Measures

We propose two kinds of error measures for performance evaluation.

The Frame based Recognition Rate (FRR) which corresponds to the percentage of correctly estimated VFOA frames, or in other words, it indicates the proportion of the time during which the VFOA has been correctly identified. However, this rate can be dominated by long duration VFOA events (where a VFOA event is defined as a temporal segment with the same VFOA label). Since we are also interested in the patterns followed by the VFOA events, which contains information related to the interaction, we also need a measure reflecting how well these events, short or long, are recognized.

Event based precision/recall, and F-measure. Let us consider two sequences of VFOA events, the GT sequence G obtained from the VFOA human annotations and the recognized sequence R obtained through the VFOA estimation process. The GT sequence is defined as $G = (G_i = (l_i, I_i = [b_i, e_i])_{i=1,\dots N_G}$ where N_G is the number of events in the ground truth G, $l_i \in \mathcal{F}$ is the ith VFOA event label, b_i and e_i the beginning and end time instants of the event l_i. The recognized sequence R is defined similarly. To compute the performance measures, the two sequences are first aligned using a string alignment procedure that takes into account the temporal extent of the events. More precisely, the matching distance between two events G_i and R_j is defined as:

$$d(G_i, R_j) = \begin{cases} 1 - F_I & \text{if } l_i = l_j \text{ and } I_\cap = I_i \cap I_j \neq \emptyset \\ 2 & \text{otherwise (i.e. events do not match)} \end{cases} \tag{19}$$

$$\text{with} \qquad F_I = \frac{2\rho_I \pi_I}{\rho_I + \pi_I}, \quad \rho_I = \frac{|I_\cap|}{|I_i|}, \quad \pi_I = \frac{|I_\cap|}{|I_j|} \tag{20}$$

where $|.|$ denotes the cardinality operator giving the size of a set. In this definition, F_I measures the degree of overlap between two events. Then, given the alignment we can compute for each person, the recall ρ_E, the precision π_E, and the F-measure F_E measuring the events recognition performances and defined as:

$$\rho_E = \frac{N_{matched}}{N_G} \quad , \quad \pi_E = \frac{N_{matched}}{N_R} \quad \text{and} \quad F_E = \frac{2\rho_E \pi_E}{\rho_E + \pi_E} \tag{21}$$

where $N_{matched}$ represents the number of events in the recognized sequence that match the same event in the GT after alignment. According to the definition in Equation 19, events are said to match

acronyms	description
gt	the head pose measurements are the ground truth data obtained with the magnetic sensor
tr	the head pose measurements are those obtained with the head tracking algorithm
gmm	the VFOA model is a GMM
hmm	the VFOA model is an HMM
ML	maximum likelihood approach: the meeting used for testing is used to train the model parameters
p	the VFOA priors (π for GMM, A for HMM) learnt from data
ge	parameters of the Gaussian were set using the geometric gaze approach
ad	VFOA model parameters were adapted

Table 3: Model acronyms: combinations of acronyms describe which experimental conditions are used. For example, gt-gmm-ge-ad specifies an adapted VFOA GMM model applied to ground truth pose data where the Gaussian parameters before adaptation were given by the geometric gaze model.

whenever their common intersection after alignment is not empty (and labels match). Thus, one may think that the counted matches can be generated by spurious accidental matches due to very small intersection. In practice, however, we observe that it is not the case and that the vast majority of matched events are consistent and have a significant degree of overlap, as illustrated in Figure 10, with 90% of the matches exhibiting a percentage of overlap higher than 50%. Even in the case of the noisier tracking data, the overlap of the correctly recognized events and their GT counterpart still match well.
In Equation 21, the recall measures the percentage of ground truth events that are correctly recognized while the precision measure the percentage of estimated events that are correct. Both precision and recall need to be high to characterize a good VFOA recognition performance. The F-measure, defined as the harmonic mean of recall and precision, reflects this requirement. Finally, the performance measures reported over the whole database (for each seating position) are the average of the precision, recall and F-measure of the 8 individuals.

7.2 Experimental protocols

To study the different modeling aspects, several experimental conditions have been defined. These conditions are summarized in Table 3 with the acronyms that will identify them in the result tables. Besides, a summary of all parameters involved in the modeling is displayed in Table 4.
First, there are two alternatives regarding the head pose measurements: the ground truth *gt* case, where the data are those obtained using the FOB magnetic field sensor, and the *tr* case which relies on the estimates obtained with the tracking system described in Section 4. In both cases, the same data origin is used for training and testing. Secondly, there are the two VFOA models, *gmm* and *hmm*, as described in Subsections 5.1 and 5.2.
Regarding learning, the default protocol is the leave-one-out approach: each meeting recording is in turn left aside for testing, while the data of the 7 other recordings are used for parameter learning, including the hyper-parameter selection in the adaptation case (denoted *ad*). The maximum likelihood case *ML* is an exception, in which the training data for a given meeting recording is composed of only the same single recording. Also, by default, the prior model parameters π or A are set to their 'uniform' values π^u or A^u, as discussed in Subsection 5.3. If these parameters are actually learned from the training data, this will be specified with a p in the result tables. Note that in the adaptation case, the hyper-parameters of the prior distribution on these parameters are always set according to the 'uniform' values. As for the *ge* acronym, it denotes the case where the VFOA Gaussian means and covariances were set according to the geometric model described in Subsection 5.4 instead of being learned from the training data. Finally, the adaptation hyper-parameter pair (s, τ) was selected (in the cartesian set $C^s \times C^\tau$) by cross-validation over the training data. The selected hyper-parameters are those that maximize the VFOA F-measure computed over the training set.

Model parameters	
μ_i, Σ_i	Gaussian parameters - learned (μ_i^l, Σ_i^l) or given by geometric modeling (μ_i^g, Σ_i^g), cf Subsection 5.3 and 5.4.
π, A	GMM and HMM model priors - learnt or set by hand to 'uniform' values π^u, A^u, cf Subsection 5.3.
$\kappa_\alpha, \kappa_\beta$	gaze factors - set by hand.
Adaptation hyper-parameters	
s	scale factor of Dirichlet distribution - set through cross-validation.
$w_i, b_{i,j}$	Dirichlet prior values of π_i and $a_{i,j}$ - set to π_i^u and $a_{i,j}^u$.
τ	scale factor of Normal prior distribution on mean - set through cross-validation.
m_i	VFOA mean prior value of Normal prior distribution - set to either μ_i^l or μ_i^g.
α	scale factor of Wishart prior distribution on covariance matrix - set by hand.
V_i	VFOA covariance matrices prior values in Wishart distribution - set to either $(\alpha-2)\Sigma_i^l$ or $(\alpha-2)\Sigma_i^g$.

Table 4: VFOA Modeling parameters: description and setting.

data	ground truth (gt)					tracking estimates (tr)				
modeling	ML	gmm	gmm-p	hmm	hmm-p	ML	gmm	gmm-p	hmm	hmm-p
FRR	79.7	72.3	74.8	72.3	72.5	57.4	47.3	51.3	47.4	48.2
recall	79.6	72.6	69.6	65.5	65.3	66.4	49.1	44.8	38.4	37.6
precision	51.2	55.1	56.2	66.7	66.5	28.9	30	39.5	59.3	60.1
F-measure F_E	62	62.4	61.9	65.8	65.6	38.2	34.8	39.3	45.2	45.3

Table 5: Average VFOA estimation results for person left under different experimental conditions (see Table 3).

8 Experiment Results

This Section describes the experiments conducted to study the behaviour of our VFOA models. We first analyze the results obtained using ground truth (GT) data, discussing the effectiveness of the modeling w.r.t. different issues (relevance of head pose to model VFOA gaze targets, predictability or stability of VFOA head pose parameters, influence of priors). In a second step, we compare the results obtained with the tracking estimates with those obtained with the ground truth, in the light of the tracking error characteristics. Then, we comment the results of the adaptation scheme, and finally, we examine more specifically the results obtained using the geometric modeling. In all cases, results are given separately for the left and right persons (see Fig. 2).

8.1 Results exploiting the GT head pose data

In this section we provide the VFOA estimation results when the head pose measurements are given by the flock-of-birds magnetic sensors.

VFOA and head pose correlation: Table 5 and 6 display the VFOA recognition results for the person left and right respectively. The first column of these two tables give the results of VFOA maximum likelihood estimation (ML) with a GMM modeling. These results show, in an optimistic case, the performances our model can achieve, and illustrate somehow the correlation between a person's head poses and his VFOA. As can be seen, this correlation is quite high for the person left (almost 80% FRR), showing the good concordance between head pose and VFOA. This correlation, however, drops to near 69% only for the right person. This can be explained by the fact that for person right, there is a strong ambiguity between looking at person left and at the slide screen, as illustrated by the empirical distributions of the pan angle in Figure 9. Indeed, the range of pan values within which the three other meeting participants and slide screen VFOA targets lies is almost half the pan range of the person left. The average angular distance between these targets is around 20 degrees for person right, a distance which can easily be covered using only eye movements rather than a head pose rotation when changing of target focus. The values of the confusion matrices, displayed in Figure 11 corroborate this analysis. The analysis of Tables 5 and 6 shows that this discrepancy holds for all experimental conditions and algorithms (when using GT head pose data), with a performance decrease from person left to person right of approximately 13% and 6% for FRR and event F-measure respectively.

VFOA Prediction: While the ML condition is achieving very good results, its performances are biased because of the mixing of training and testing data. On the contrary, the GMM and HMM modelling

data	ground truth (gt)					tracking estimates (tr)				
modeling	ML	gmm	gmm-p	hmm	hmm-p	ML	gmm	gmm-p	hmm	hmm-p
FRR	68.9	56.8	61.6	57.3	61.6	43.6	38.1	49.1	38	38.3
recall	72.9	66.6	65.1	58.4	58.2	65.6	55.9	48.7	37.3	37.4
precision	47.4	49.9	51.4	63.5	64.1	24.1	26.8	35.2	55.1	55.9
F-measure F_E	56.9	54.4	55.8	59.5	59.7	34.8	35.6	40.4	43.8	44.2

Table 6: VFOA estimation results for person right under different experimental conditions (see Table 3).

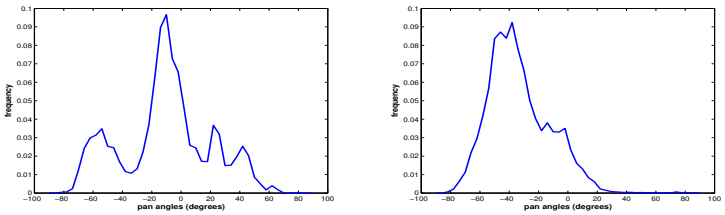


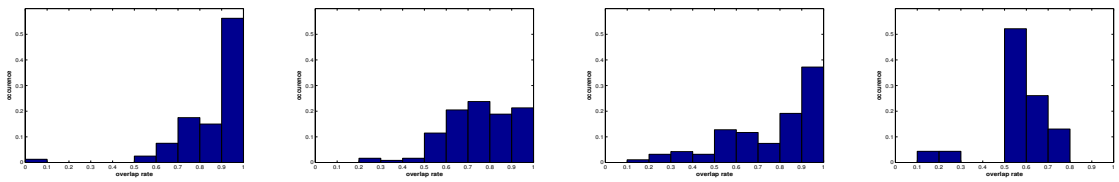
Figure 9: Empirical distribution of the GT head pose pan angle computed over the database for person left (left image) and right. For person left, the people and slide screen VFOA targets can still be identified through the pan modes. For person right, the degree of overlap is quite significant.

are showing the generalization property of the modelling, by learning the VFOA parameters from other persons' data. From the Table 5 and 6, we observe that the GMM and HMM modelling with or without a prior term produce results close to the ML case. For both person left and right, the GMM approach is achieving better performances in term of frame recognition rate and event recall while the HMM is giving better event precision and F measure. This can be explained since the HMM approach is mainly doing data smoothing. As a results some events are missed (lower recall) but the precision increases due to the elimination of short spurious detections.

VFOA Confusions: Figure 11 a) and b) display as images the confusion matrices obtained with the VFOA frame recognition performance measure and an HMM modelling. The confusion matrices for both person left and right with GT head pose data clearly exhibit confusion between near VFOA targets. For instance, for person left, $O2$ is sometimes confused with PR or $O1$. For person right, the main source of confusion is between PL and SS, as already mentioned. In both cases, the table T can be confused with $O1$ and $O2$, as can be expected since these targets share more or less the same pan values. Thus, most of the confusion can be explained by the geometry of the room and the fact that people can modify their gaze without modifying their head pose, and therefore do not always need to turn their head to focus on a specific VFOA target.

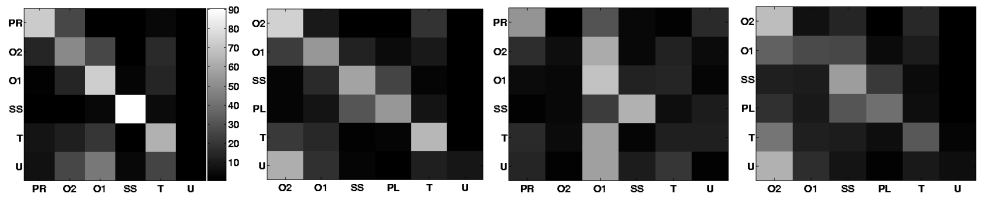
Influence of Priors: Table 5 and 6 also present the recognition rates when learning the prior on the events (-p extension). As can be seen, while the improvement is moderate using the GT head pose data or the HMM modeling, it is quite beneficial in the GMM case when working with the tracking pose estimates. The effect of the prior is illustrated in Figure 12. While the $O2$ VFOA has its decision area reduced, $O1$ sees its decision surface extended because its VFOA event is more represented in our database. In practice, the VFOA distribution prior allows to clearly favor most likely events while almost removing less likely events in some extreme cases. Although results show that taking priors into account can significantly improve the performance, their usage could clearly be a problem when using the VFOA recognition system on other meetings with different VFOA structures, or if the same people (e.g. $O1$ and $O2$) would have exchanged their seats across meetings. Thus, in the remaining of the result analysis, we will not use such prior in the experiments.

Comparison with other algorithms: We can compare our VFOA recognition performances to other state of the art VFOA estimation algorithms based on GT head pose data. [12] have conducted an interesting work about VFOA interaction analysis, where one of the task, among others, consisted



(GT, Left) (GT, Right) (TR, Left) (TR, Right)

Figure 10: Distribution of overlap measures F_I between true and estimated matched events. The estimated events were obtained using the HMM approach. GT and TR respectively denote the use of GT head pose data and tracking estimates data. Left and Right denote person left and right respectively.



(a) (GT, Left) (b) (GT, Right) (c) (TR, Left) (d) (TR, Right)

Figure 11: Frame-based recognition confusion matrices obtained with the HMM modeling (gt-hmm and tr-hmm conditions). VFOA targets 1 to 4 have been ranked according to their pan proximity: PR: person right - PL: person left - O1 and O2: organizer 1 and 2 - SS: slide screen - T: table - U: unfocused.

in estimating the VFOA of four people engaged in a conversation, using people's speaking status and head pose measured with magnetic field sensors. For each person, the potential VFOA were the three other participants. They obtained an average frame based recognition rate of 67.9 %. Despite the lower number of VFOA targets and the multiple cues they were using (speech and magnetic sensors output), their results are similar to ours. We obtained 57% for person right and 72.3% for person left using the HMM recognizer (resp. 62% and 72.7% with adaptation, as shown later).

8.2 Results with Head Pose Estimates

Table 5 and 6 provide the VFOA recognition performance obtained using the head pose tracking estimates, under the same experimental conditions than when using the GT head pose data. As can be seen, significant performance degradation can be noticed. In the ML case, the decrease in FRR and F-measure ranges from 22% to 26% for both person left and right. These degradations are mainly due to tracking errors of different types: small pose estimation errors, and also sometimes large errors due to short periods when the tracker locks on a subpart of the face. Figure 12 illustrates the effect of the pose estimation errors, and in particular of the tilt ones, on the VFOA distributions. While the increase of VFOA pan variances is moderate when moving from GT head pose data (first row) to pose estimates (second row), it is quite important in the tilt direction, as can be observed on the VFOA decision maps.

When analyzing in more details the results of Table 5 and 6, one can notice that while the performance decrease using the GMM follows the ML case, the deterioration for the HMM is smaller, in particular when considering the F-measure. This demonstrates that, whereas with the GT head pose data the HMM modelling did not have much impact on performances w.r.t. the GMM, in presence of noisier data, the HMM smoothing effect is quite beneficial. Also, the HMM performance decrease is smaller

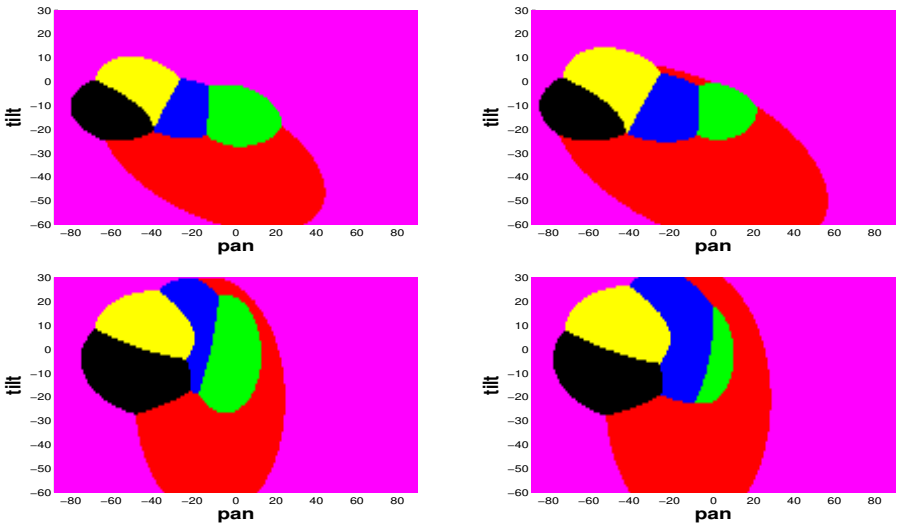


Figure 12: Pan-tilt space VFOA decision maps for person right obtained from all meetings, in the GMM case (cf Eq. 4), working with GT head pose data (first row) or tracking estimates (second row), and by learning (second column) or not (first column) the VFOA priors. black=PL, yellow=SS, blue=$O1$, green=$O2$, red=T, magenta =U.

error measure	gt-gmm	gt-gmm-ad	gt-hmm	gt-hmm-ad	tr-gmm	tr-gmm-ad	tr-hmm	tr-hmm-ad
FRR	72.3	72.3	72.3	72.7	47.3	57.1	47.4	53.1
recall	72.6	72.1	65.5	68.8	49.1	48.7	38.4	40.5
precision	55.1	53.7	66.7	64.4	30	41	59.3	62.5
F-measure F_E	62.4	61.2	65.8	66.2	34.8	42.8	45.2	47.9

Table 7: Average VFOA estimation results for person left, before and after adaptation.

for person right (19% and 15% for respectively the FRR and F measure) than for person left (25% and 20%). This can be explained by the better tracking performance -in particular regarding the pan angle- achieved on people seated at the person right position (cf Table 1). Figure 13 presents the plot of the VFOA FRR versus the pan angle tracking error for each meeting participant, when using GT head pose data (in this case the tracking error is 0) or pose estimates. It shows that for left people, there is a strong correlation between tracking error and VFOA performance, which can be explained by the fact that higher tracking errors directly generates larger overlap between VFOA class-conditional pose distributions (cf Fig. 9, left). For right people, this correlation is weaker, as good tracking can result in bad VFOA recognition performance. In this case, the higher level of inherent ambiguities between several VFOA targets (e.g. SS and PL) may play a larger role.
Finally, the 2 right images of Fig. 11 display the confusion matrices when using the HMM model and the head pose estimates. The same confusion than using the GT head pose data are exhibited, but more pronounced because of the tracking errors (see above) and tilt estimation uncertainties.

8.3 Results with Model Adaptation

Tables 7 and 8 display the recognition performance obtained when using the adaptation framework described in Section 6[5]. When considering the left person position, one can observe no improvement when using GT head pose data and a large improvement when using the tracking estimates (e.g. around 10% and 8% for resp. FRR and F_E with the GMM model). In this situation, the adaptation

[5]We recall the values without adaptation for ease of comparison.

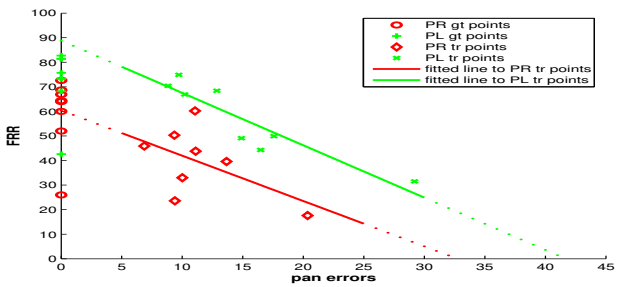


Figure 13: VFOA frame based recognition rate vs head pose tracking errors (for the pan angle), plotted per meeting. The VFOA recognizer is the HMM modeling after adaptation.

error measure	gt-gmm	gt-gmm-ad	gt-hmm	gt-hmm-ad	tr-gmm	tr-gmm-ad	tr-hmm	tr-hmm-ad
FRR	56.8	59.3	57.3	62	38.1	39.3	38	41.8
recall	66.6	70.2	58.4	63	55.9	55.3	37.3	43.6
precision	49.7	50.9	63.5	64.5	26.8	29	55.1	56.1
F-measure F_E	54.4	56.4	59.5	62.7	35.6	37.3	43.8	48.8

Table 8: Average VFOA estimation results for person right, before and after adaptation.

is able to cope with the tracking errors and the possible variability among people in head pose response to different appearances. For person right, we notice improvement with both the GT and tracking head pose data. For instance, with the HMM model and tracking data, the improvement is 3.8% and 5% for FRR and F_E. Again, in this situation adaptation can cope for people's personal way of looking to the targets, such as correcting the bias in head tilt estimation, as illustrated in Figure 14.
When exploring the optimal adaptation parameters estimated through cross-validation, one obtain the histograms of Figure 15. As can be seen, regardless of the input pose data, they correspond to configurations giving approximately equal balance to the data and prior regarding the adaptation of the HMM transition matrices (s_1 and s_2), and configurations for which the data are driving the adaptation process of the mean pose values (τ_1 and τ_2).

Comparison with other algorithms: Our results, 42% FRR for person right and 53% for person left, are quite far from the 73% reported in the interesting paper of [11]. Several factors may explain the difference. First, for [11], a 4 people meeting situation was considered and no other VFOA target apart from the other meeting participants was considered. In addition, these participants were sitting at equally spaced angles around a round table, optimizing the discrimination between VFOA targets. From a tracking point of view, people were recorded from a camera placed in front of them. Thus, due to the table geometry, the very large majority of head pan lay between $[-45, 45]$ degrees, where the tracking errors are smaller (see Table 2) [6]. Ultimately, our results are more in accordance with the 52% FRR reported by the same authors [41] when using the same framework [11] but applied to a 5-person meeting, resulting in 4 VFOA targets.

8.4 Results with a the Geometrical VFOA Modelling

In this Section we study the approach based on the models exploiting the geometry of the meeting room, as described in Subsection 5.4. The possibility to set the VFOA parameters (mean, covariances) from geometry is interesting because it may remove the need for data annotation each time a new VFOA target is considered, e.g. when people are moving around in the room.
Figure 17 shows the geometric VFOA head Gaussian parameters (mean and covariance) generated by

[6] Furthermore, it seems from the paper that the head pose tracking algorithm was trained on the face images of the same people appearing in the test video, which would result in even smaller tracking errors

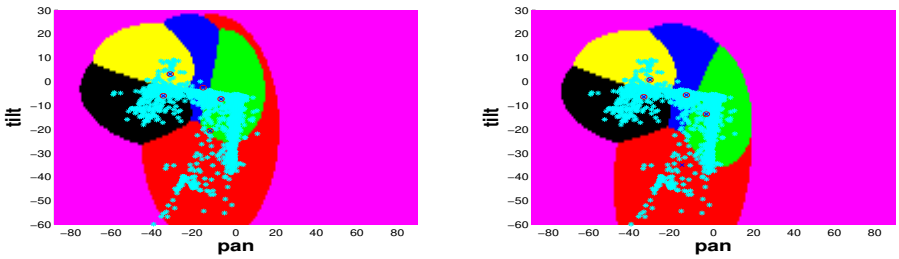


Figure 14: VFOA decision map example before adaptation (Left) and after adaptation (right). After adaptation, the VFOA of $O1$ and $O2$ correspond to lower tilt values. black=PL, yellow=SS, blue=$O1$, green=$O2$, red=T, magenta =U. The blue stars represent the tracking head pose estimates used for adaptation.

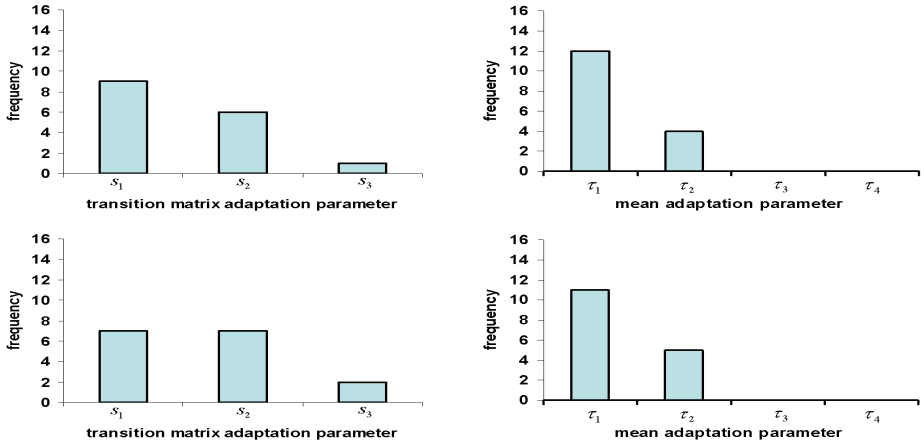


Figure 15: Histogram of the optimal scale adaptation factor of the HMM prior (first column) and HMM VFOA mean (second column), selected though cross-validation on the training set, and when working with GT head pose data (first row) or with tracking head pose estimates (second row).

the model when using $(\kappa_\alpha, \kappa_\beta) = (0.5, 0.4)$. As can be seen, the VFOA pose values predicted by the model are consistent with the average pose values computed for individuals using the pose GT head pose data, specially for the person left position. For the person right's position, we can observe that the geometric model is wrongly predicting the gaze value of $= O2$ (not moved) and $O1$ (attraction in the wrong direction). Indeed, for person right, our assumption that the rest head orientation N in Fig. 6 consists of looking on the other side of the table is not appropriate: as all the VFOA targets are located on their right side, people tend to already orient their shoulder towards their right as well (see Fig 16), and thus N should be set accordingly. Assuming that the rest looking direction corresponds to looking at $O1$, we obtain a better match. This is demonstrated by Table 9, which provides the prediction errors in pan E_{pan} defined as:

$$E_{pan} = \frac{1}{8 \times (K-1)} \sum_{m=1}^{8} \sum_{f_i \in \mathcal{F}/\{U\}} |\bar{\alpha}_m(f_i) - \alpha_m^p(f_i)| \tag{22}$$

where $\bar{\alpha}_m(f_i)$ is the average pan value of the person in meeting m and for the VFOA f_i, and $\alpha_m^p(f_i)$ is the predicted value according to the chosen model (i.e. the pan component of $\mu_{f_i}^g$ or $\mu_{f_i}^l$ in the

method	learned VFOA		geometric VFOA	
error	E_{pan}	E_{tilt}	E_{pan}	E_{tilt}
L	6.37	5.08	5.54	6.35
R (ref:looking straight)	5.85	6.07	12.5	7.65
R (ref: looking at O1)	5.85	6.07	5.62	7.65

Table 9: Prediction errors for learned VFOA and geometric VFOA models when using GT head pose data(R for right and L for person left). For person right ref1 correspond to rest head orientation=looking straight, ref2= rest head orientation=looking at $O1$

Figure 16: Rest direction for person right: the person right turn himself toward O1 instead of looking straight in front of him (toward $O2$)

geometric or learning approaches respectively). The tilt prediction error E_{tilt} is obtained by replacing in Equation 22 pan angles by tilt angles.

The VFOA recognition performances with the geometrical modelling are presented in Tables 10 and 11. For person right, the model using as rest head pose looking at $O1$ is used. These tables show that, when using GT head pose data, the results are worse than with the learning approach, which is somewhat surprising given the similarity in the prediction errors. Fortunately, with the head pose tracking data, the results are similar. Given that the modeling does not request any training data (except for camera calibration), this is an interesting result. Also, we can notice that adaptation also improves the recognition, though only for the person left.

9 Conclusion and Future Work

In this paper we presented a methodology to recognize the VFOA of meeting participants from their head pose, the latter being defined by its pan and tilt angles. Such head pose measurements were obtained either through magnetic field sensors or using a probabilistic based head pose tracking algorithm. The experiments showed that, depending on people's position in the meeting room and on the angular distribution of the VFOA targets, *the eye gaze may or may not be highly correlated with the head pose.*

In absence of such correlation, and if eye white/gaze tracking is unaccessible due to low resolution images, the only way to improve VFOA recognition may only come from the prior knowledge embedded in the cognitive and interactive aspects of human-to-human communication. Ambiguous situations such as the one illustrated in Figure 18, where the same head pose can correspond to two different VFOA targets, could be resolved by the joint modeling of the speaking and VFOA characteristics of all meeting participants. Such characteristics have been shown to exhibit specific patterns/statistics in the behavioral and cognitive literature, as already exploited by [12]. This will be the topic of future research.

Besides, as shown by the experiments, there indeed exists some correlation between head pose tracking errors and VFOA recognition results. Improving the tracking algorithms, e.g. using multiple cameras, higher resolution images or adaptive appearance modeling techniques, would thus improve the VFOA results. Finally, in the case of meetings in which people are moving to the slide screen or white board for presentations, the development of a more general approach that models the VFOA of these moving

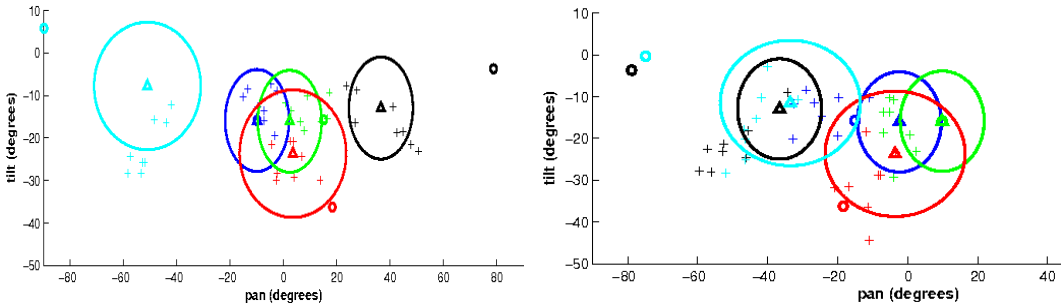


Figure 17: Geometric VFOA Gaussian distributions for person left (left image) and person right (right): the figure displays the gaze target direction (◯), the corresponding head pose contribution according to the geometric model with values $(\kappa_\alpha, \kappa_{tilt}) = (0.5, 0.4)$ (△ symbols), and the average head pose (from GT head pose data) of individual people (+). Ellipses display the standard deviations used in the geometric modelling. black=PL or PR, cyan=SS, blue=$O1$, green=$O2$, red=T.

measure	gt	gt-ge	gt-ad	gt-ge-ad	tr	tr-ge	tr-ad	tr-ge-ad
FRR	72.3	65.8	72.3	70.5	47.4	45.2	53.1	52.2
recall	72.1	65.3	68.8	67.4	38.4	49.1	40.5	48.9
precision	55.1	49.5	64.4	55.2	59.3	41.1	62.5	46.7
F-measure F_E	61.2	56.2	66.6	60.4	45.2	43.6	47.9	46.6

Table 10: Average VFOA estimation results for person left using the HMM model with geometric VFOA parameter setting, with/without adaptation and $(\kappa_\alpha, \kappa_{tilt}) = (0.5, 0.4)$.

people will be necessary. This has been one topic of our recent research [1].

References

[1] K. Smith, S.O. Ba, D. Gatica-Perez, and J.-M. Odobez, "Multi-person wandering focus of attention tracking," in *International Conference on Multimodal Interfaces*, 2006.

[2] Will Thoretz, "Press release: Nielsen to test electronic ratings service for outdoor advertising," 2002.

[3] Elizabeth M. Tucker, "The power of posters," Tech. Rep., University of Texas at Austin, 1999.

[4] J.E. McGrath, *Groups: Interaction and Performance*, Prentice-Hall, 1984.

[5] D. Heylen, "Challenges ahead head movements and other social acts in conversation," in *The Joint Symposium on Virtual Social Agent*, 2005.

[6] S.R.H. Langton, R.J. Watt, and V. Bruce, "Do the eyes have it ? cues to the direction of social attention," *Trends in Cognitive Sciences*, vol. 4(2), pp. 50–58, 2000.

[7] N. Jovanovic and H.J.A. Op den Akker, "Towards automatic addressee identification in multi-party dialogues," in *5th SIGdial Workshop on Discourse and Dialogue*, 2004.

[8] S. Duncan Jr, "Some signals and rules for taking speaking turns in conversations," *Journal of Personality and Social Psychology*, vol. 23(2), pp. 283–292, 1972.

[9] D. Novick, B. Hansen, and K. Ward, "Coordinating turn taking with gaze," in *International Conference on Spoken Language Processing*, 1996.

[10] D. Zhang, D. Gatica-Perez, S. Bengio, I. McCowan, and G. Lathoud, "Modeling individual and group action in meetings: a two-layer hmm framework," in *IEEE CVPR Workshop on Event Mining in Video*, 2004.

measure	gt	gt-ge	gt-ad	gt-ge-ad	tr	tr-ge	tr-ad	tr-ge-ad
FRR	57.3	48.5	62	48.8	38	40.6	41.8	41.9
recall	58.4	48.7	63	54.2	37.3	53.2	43.6	55
precision	63.5	56.4	64.5	52.8	55.1	43	56.1	40.8
F-measure F_E	59.5	51.2	62.7	52.2	43.8	47.3	48.8	46.4

Table 11: Average VFOA estimation results for person right using the HMM model with geometric VFOA parameter setting, with/without adaptation, and $(\kappa_\alpha, \kappa_{tilt}) = (0.5, 0.4)$.

Figure 18: Ambiguity in focus: despite the high visual similarity of the head pose of the right person, the two focus are different (left image: person left: right image: slide screen). Resolving such cases can only be done by using context (speaking status, other's people gaze, slide activity etc).

[11] R. Stiefelhagen, J. Yang, and A. Waibel, "Modeling focus of attention for meeting indexing based on multiple cues," *IEEE Transactions on Neural Networks*, vol. 13(4), pp. 928–938, 2002.

[12] K. Otsuka, Y. Takemae, J. Yamato, and H. Murase, "A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances," in *International Conference on Multimodal Interface (ICMI'05)*, 2005, pp. 191–198.

[13] ICPR-POINTING, "Icpr: Pointing'04: Visual observation of deictic gestures workshop," 2004.

[14] CLEAR, "ClEAR evaluation campaign and workshop," 2006.

[15] R. Stiefelhagen and J. Zhu, "Head orientation and gaze direction in meetings," in *Conference on Human Factors in Computing Systems*, 2002.

[16] Edward G. Freedman and David L. Sparks, "Eye-head coordination during head-unrestrained gaze shifts in rhesus monkeys," *Journal of Neurophysiology*, vol. 77, pp. 2328–2348, 1997.

[17] I.V. Malinov, J. Epelboim, A.N. Herst, and R.M. Steinman, "Characteristics of saccades and vergence in two kinds of sequential looking tasks," *Vision Research*, 2000.

[18] S. O. Ba and Jean Marc Odobez, "A rao-blackwellized mixed state particle filter for head pose tracking," in *ICMI Workshop on Multi-modal Multi-party Meeting Processing, Trento Italy*, 2005, pp. 9–16.

[19] R.G.M. Pieters, E. Rosbergen, and M. Hartog, "Visual attention to advertising: The impact of motivation and repetition," in *Conference on Advances in Consumer Research*, 1995.

[20] P. Smith, M. Shah, and N. Da Vitoria Lobo, "Determining driver visual attention with one camera," *IEEE Transaction on Intelligent Transportation Systems*, vol. 4.(4), pp. 205–218, 2004.

[21] Y. Matsumoto, T. Ogasawara, and A. Zelinsky, "Behavior recognition based on head pose and gaze direction measurement," in *Conference on Intelligent Robots and Systems*, 2002.

[22] A.H. Gee and R. Cipolla, "Estimating gaze from a single view of a face," in *International Conference on Pattern Recognition*, 1994.

[23] T. Horprasert, Y. Yacoob, and L. Davis, "Computing 3d head orientation from a monocular image sequence," in *IEEE International Conference on Automatic Face and Gesture Recognition*, 1996.

[24] R. Stiefelhagen, J. Yang, and A. Waibel, "A model-based gaze tracking system," in *IEEE International Joint Symposia on intelligence and Systems*, 1996.

[25] R. Zhang and Z. Zhang, "Model-based head pose tracking with stereo-vision," Tech. Rep. MSR-TR-2001-102, Microsoft Research, 2001.

[26] R. Rae and H. Ritter, "Recognition of human head orientation based on artificial neural networks," *IEEE Transaction on Neural Network*, vol. 9(2), pp. 257–265, 1998.

[27] V. Kruger, S. Bruns, and G. Sommer, "Efficient head pose estimation with Gabor wavelet networks," in *British Machine Vision Conference*, 2000.

[28] L. Zhao, G. Pingali, and I. Carlbom, "Real-time head orientation estimation using neural networks," in *International Conference on Image Processing*, 2002.

[29] T.F. Cootes, G. J. Edwards, and C.J. Taylor, "Active appearance models," in *European Conference on Computer Vision*, 1998, pp. 183–191.

[30] S. Srinivasan and K. L. Boyer, "Head pose estimation using view based eigenspaces," in *International Conference on Pattern Recognition*, 2002.

[31] Y. Wu and K. Toyama, "Wide range illumination insensitive head orientation estimation," in *IEEE Conference on Automatic Face and Gesture Recognition*, 2001.

[32] L. Brown and Y. Tian, "A study of coarse head pose estimation," in *IEEE Workshop on Motion and Video Computing*, 2002.

[33] R. Stiefelhagen, "Estimating head pose with neural networks-Results on the pointing04 icpr workshop evaluation data," in *Pointing 04 ICPR Workshop*, 2004.

[34] M. Danninger, R. Vertegaal, D.P. Siewiorek, and A. Mamuji, "Using social geometry to manage interruptions and co-worker attention in office environments," in *Conference on Graphics Interfaces*, 2005.

[35] M. Hayhoe and D. Ballard, "Eye movements in natural behavior," *TRENDS in Cognitive Sciences*, vol. 9(4), pp. 188–194, 2005.

[36] S. Baron-Cohen, "How to build a baby that can read minds: cognitive mechanisms in mindreading," *Cahier de psychologies Cognitive*, vol. 13, pp. 513–552, 1994.

[37] J.-M. Odobez, "Focus of attention coding guidelines," Tech. Rep. IDIAP-COM-2, IDIAP Reasearch Institute, 2006.

[38] N. Gourier, D. Hall, and J. L. Crowley, "Estimating face orientation from robust detection of salient facial features," in *Pointing 2004, ICPR International Workshop on Visual Observation of Deictic Gestures*, 2004, pp. 183–191.

[39] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Readings in Speech Recognition*, vol. 53A(3), pp. 267–296, 1990.

[40] J.L. Gauvain and C. H. Lee, "Bayesian learning for hidden Markov model with Gaussian mixture state observation densities," *Speech Communication*, vol. 11, pp. 205–213, 1992.

[41] R. Stiefelhagen, *Tracking and Modeling focus of attention*, Ph.D. thesis, University of Karlsruhe, 2002.